Evolving Classifiers to Recognise the Movement Characteristics of Parkinson's Disease Patients

Michael A. Lones, *Senior Member, IEEE,* Stephen L. Smith, *Member, IEEE,* Jane E. Alty, Stuart E. Lacy, Katherine L. Possin, D. R. Stuart Jamieson and Andy M. Tyrrell, *Senior Member, IEEE*

***Abstract*—Parkinson's disease is a debilitating neurological condition that affects approximately 1 in 500 people and often leads to severe disability. To improve clinical care, better assessment tools are needed that increase the accuracy of differential diagnosis and disease monitoring. In this paper, we report how we have used evolutionary algorithms to induce classifiers capable of recognising the movement characteristics of Parkinson's disease patients. These diagnostically-relevant patterns of movement are known to occur over multiple time scales. To capture this, we used two different classifer architectures: sliding-window genetic programming classifiers, which model over-represented local patterns that occur within time series data, and artificial biochemical networks, computational dynamical systems that respond to dynamical patterns occurring over longer time scales. Classifiers were trained and validated using movement recordings of 49 patients and 41 age-matched controls collected during a recent clinical study. By combining classifiers with diverse behaviours, we were able to construct classifier ensembles with diagnostic accuracies in the region of 95%, comparable to the accuracies achieved by expert clinicians. Further analysis indicated a number of features of diagnostic relevance, including the differential effect of handedness and the over-representation of certain patterns of acceleration.**

***Index Terms*—Automated disease diagnosis, Time series analysis, Classification, Genetic programming, Artificial biochemical networks.**

I. Introduction

PARKINSON'S disease (PD) is a chronic, progressive, neurodegenerative disorder with a characteristic motor syndrome caused by a loss of dopaminergic neurons in the brain. While genetic and environmental factors (e.g., pesticide exposure) have been shown to increase the risk for developing PD, in most cases the cause is unknown [1]. It is one of the most common neurodegenerative disorders, typically developing between 50–70 years of age. Parkinson's UK estimates that there are a total of four million people with PD worldwide, and that 1 in every 500 people in the UK has PD [2]. Within countries with ageing populations, such as the USA and most of Europe, it is expected that the number of cases of PD will triple in the next 50 years [3].

Although there is currently no cure for PD, early and suitable treatment greatly increases quality of life [4]. Misdiagnosis rates are high, with estimates ranging from 8% to as high as 25% [5]–[7]. This is because the clinical presentation of idiopathic PD can mimic other neurological conditions including essential tremor, progressive supranuclear palsy, multiple system atrophy, corticobasal syndrome, and vascular parkinsonism [8]. Even when accurately diagnosed, making optimal treatment decisions can be daunting for the practising clinician because of the number of drug and dosing options available, the variability in how different patients respond to the same medications, and the variability in how an individual patient's treatment response will change throughout the course of the disease [9]. Incorrect medication selection and dosage can lead to unpleasant, difficult to treat side-effects, such as levodopa-induced dyskinesia and hallucinations.

Clinicians fine-tune a patient's medication regimen based on their own ratings of the patient's symptoms (e.g. using the Unified Parkinson's Disease Rating Scale [10], which rates symptoms on 5-point scales) and on patient-rated treatment response. These metrics may be insensitive to small but important effects, and of concern, these metrics correlate only weakly with each other [11]. Better tools are needed to support clinicians in making accurate diagnoses and in monitoring medication regimens. The purpose of this study was to investigate the feasibility of evolved classifiers applied to movement data for detecting symptoms[1] of PD.

Although the symptoms of PD are variable, all patients experience some form of movement disorder, including slowing of movement, tremor, rigidity and impaired balance. Bradykinesia is the diagnostically most-relevant symptom of PD; literally 'slow movement', general use of the term also encompasses delays or hesitations in movement, sparsity of movement, and poor rhythmic control. Bradykinesia is clinically assessed using rapidly alternating movements, such as finger-tapping, where the patient is asked to repeatedly tap together their thumb and forefinger. Ratings are based on

This research was supported in part by the White Rose University Consortium and the EPSRC under the grant "Artificial Biochemical Networks: Computational Models and Architectures" (ref. EP/F060041/1).

Michael Lones is with the School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, U.K., having previously been at the University of York. (e-mail: michael.lones@york.ac.uk).

Stephen Smith, Stuart Lacy and Andy Tyrrell are with the Intelligent Systems Group, Department of Electronics, University of York, U.K. (e-mail: stephen.smith@york.ac.uk, sl561@york.ac.uk, andy.tyrrell@york.ac.uk).

Jane Alty and Stuart Jamieson are with the Department of Neurology at Leeds General Infirmary, Leeds, U.K. (e-mail: altyjane@doctors.org.uk, Stuart.Jamieson@leedsth.nhs.uk).

Katherine Possin is with the Memory and Aging Center, University of California, San Francisco, U.S.A. (e-mail: Katherine.Possin@ucsf.edu).

[1]Or more correctly, in medical terms, the *signs* of PD.

the clinician's perceived abnormality of this movement. Rest tremor is also characteristic of PD, and is typically evaluated by clinician observation while the patient's limbs are at rest. Even for highly-trained clinicians, there is considerable inter-rater and intra-rater inconsistency in judging the severity of these cardinal symptoms [12], [13] which impairs both diagnosis and monitoring of PD.

We have previously discussed the possibility of developing a non-invasive computer-based assessment for PD, which would objectively measure a patient's movements [14]–[16]. In a small feasibility study, the movements of 12 PD patients and 10 age-matched control subjects were recorded whilst they traced a geometric design using a graphics tablet. We found that both genetic programming (GP) [15] and artificial immune systems [16] were able to classify the PD patients by recognising over-represented patterns of movement. However, the small sample sizes made it difficult to ascertain the generality of these results with respect to the wider normal and disease populations.

Other researchers have also applied machine learning algorithms to PD classification, many reporting accuracies in excess of 90% [17], [18]. However, the use of small samples (particularly of non-PD subjects) again makes it hard to determine the generality of these figures. For instance, Tsanas et al. [18] report an accuracy of 99% when using support vector machines to discriminate between vocal recordings of 33 PD patients and 10 controls.

In addition to supporting clinicians in identifying and accurately measuring parkinsonian motor symptoms, techniques such as GP have the potential to support the discovery of novel information about symptoms of the underlying disease. For instance, in related work on cancer diagnosis [19] and the evaluation of visuo-spatial ability [20], we carried out analysis of evolved classifier populations to identify conserved patterns within the data. This kind of approach could be particularly relevant to PD, where understanding of the disease's causes, symptoms and subtypes is incomplete.

In this paper, we report the results of a much larger clinical study in which movement data was collected from 49 PD patients and 41 age-matched controls as they performed a variety of tasks. Rather than the geometrical figure tracing task used in our earlier work, we used an electromagnetic motion capture device to record subjects' movements whilst performing standard PD clinical assessment tasks. This has the advantage of maintaining the existing testing environment and its associated metrics. Using this data, we evolved two types of programmatic classifiers to discriminate between subjects with and without PD. Analysis of their behaviour indicated their discriminative abilities to be based on recognition of a number of different patterns within the movement data. By combining behaviourally diverse classifiers of each type, we were able to construct ensembles that were highly accurate in detecting PD motor symptoms. Results suggest that the application of evolved classifiers to automated movement data is a promising method for the development of new diagnostic and monitoring tools.

The paper is organised as follows: Section II provides a summary of the movement data used to train and validate classifiers. Section III introduces the classifier architectures and performance measures used in this work. Section IV presents baseline metrics, details of evolved classifiers, behavioural analysis, and formation of classifier ensembles. Section V discusses the clinical interpretation of these results and the implications for biomedical data mining more generally. Section VI concludes.

II. Clinical Study Data

A. Subjects

Test subjects were recruited from clinics held at the Leeds Teaching Hospitals NHS Trust, UK, between August 2009 and October 2010. Forty-nine Parkinson's patients participated in the study, each previously diagnosed by a neurologist. Forty-one age-matched controls were recruited from patients' spouses and companions and staff in the neurology department. Mean ages were 67 years ($\pm$9) for patients and 64 years ($\pm$10) for controls. Male to female ratios were 31:18 for patients and 14:27 for controls, reflecting the higher incidence of PD in men [21]. Right to left-handedness ratios were 41:8 for patients and 33:8 for controls. The study was granted approval by the National Research Ethics Service and Medicines and Healthcare Products Regulatory Agency. Written informed consent was obtained from all subjects, and their medications were not altered for the study[2]. There was no history of neurological disease amongst the control subjects.

B. Movement Tasks

Movement data was collected using a Polhemus Patriot electromagnetic motion tracking device, whose probes were attached to the subject's thumb and index finger whilst carrying out prescribed tasks. The Polhemus Patriot has a sampling rate of 60Hz, and measures both position and orientation relative to a point source in real-time.

1) Finger Tapping: Finger tapping is a standard clinical test for assessing bradykinesia. The subject was asked to tap their thumb and index finger repeatedly for a duration of 30 seconds, using each hand in turn. Subjects were asked to carry out this exercise as rapidly as possible, separating the finger and thumb as far as they could comfortably achieve.

2) Movement at Rest: Tremor is commonly seen in PD patients, and typically occurs when a subject is at rest. Whilst still connected to the Polhemus Patriot, the subject was asked to place their hand in a resting position on the arm of a chair. They were then asked to count backwards from 100 in order to distract them from consciously correcting any involuntary movement. Motion data was recorded for a duration of 30 seconds for each hand.

3) UPDRS Ratings: During both tasks, a neurologist with speciality training in movement disorders performed the UPDRS 'tremor at rest' and 'finger taps' components, which are scored between 0 and 4 [10]. For tremor at rest, 0 corresponds to no tremor and 4 corresponds to tremor that is marked in

[2]A requirement for ethical consent. Since medication reduces symptoms, this makes it harder to discriminate between patients and controls, and consequently makes diagnosis more difficult. However, the task is comparable to clinical monitoring, where patients are already undergoing drug treatment.

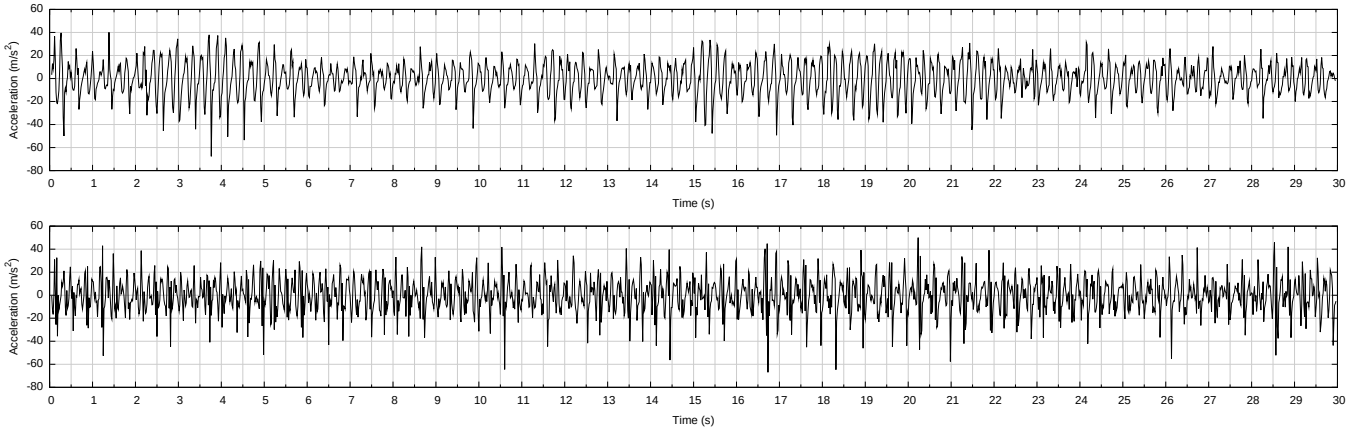


Fig. 1. Example recordings of two PD patients carrying out finger tapping.

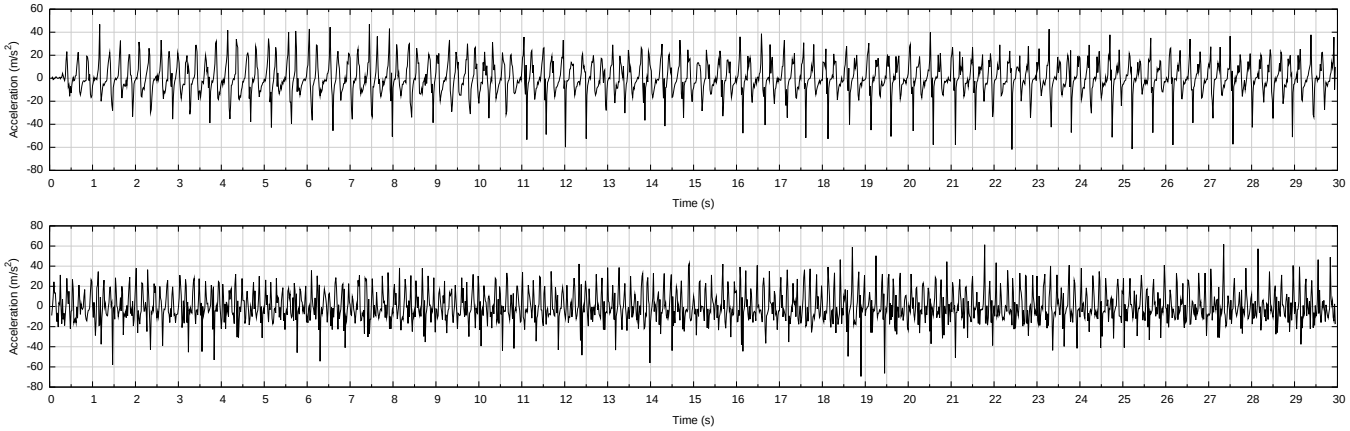


Fig. 2. Example recordings of two age-matched controls carrying out finger tapping.

amplitude and is present most of the time. For finger tapping, 0 corresponds to normal and 4 indicates that the task can barely be performed.

III. Classification

A classifier is a mapping from a set of data to a set of labelled classes. Classifiers are induced by training on a subset of data for which class membership is already known. Following training, the induced classifier can then be used to predict the class membership of data which was not seen during training. Evolutionary algorithms are a widely-used method for inducing classifiers [22]–[25]. Factors which make them effective for classification problems include their breadth of search, relatively low sensitivity to initial conditions, and flexibility in terms of representation and evaluation of solutions [26]. They are particularly useful for problems where there is limited prior understanding of what a solution should look like, where the method's breadth of search and ability to use relatively unconstrained solution representations permits a wide exploration of candidate solutions. This includes many problems in biology [27] and medicine [28], where classification often involves modelling processes that are complex, dynamical and poorly understood. Examples include the discovery of relationships in human genetics [29], the interpretation of noisy biochemical spectral data [19], and the modelling of genetic sequences [30]. The latter domain offers relevant examples of where evolutionary algorithms have been used to evolve relatively unusual classifier architectures, for instance the use of programmatic expressions [31] and augmented state machines [32] to describe conserved patterns of DNA bases. In this regard, evolutionary algorithms appear well suited to neurological diagnosis, a domain in which there is often limited understanding of the underlying biology of the diseases, and consequently limited understanding of the most appropriate classifier models to use.

A. Classifying Finger Tapping Recordings

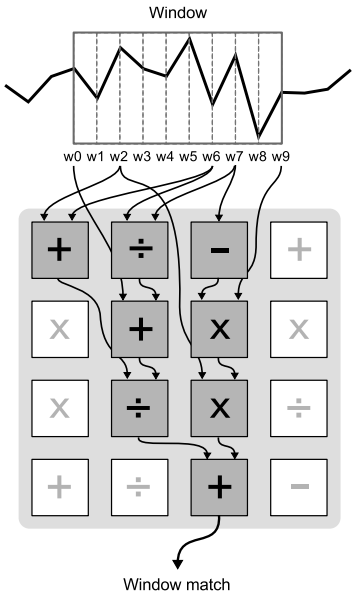
There is considerable variation in the way people carry out finger tapping. PD is one factor that affects how people tap their fingers, but other factors could include poor dexterity caused by age, arthritis, or other pathological conditions. As such, it can be difficult to discriminate PD and control recordings through visual inspection. This is reflected in Figures 1 and 2, which show example recordings of finger tapping carried out by PD patients and age-matched controls. Nevertheless, we know that PD affects the way in which people move, and can therefore assume that there are characteristic patterns of movement which can be used to discriminate PD

patients from normal controls. We can also assume there are multiple patterns, which occur over multiple time-scales. For instance, during individual taps, we can expect to see the cog-wheel-like motion associated with PD movement [33]. Over longer time-scales, we may see patterns of change in amplitude, frequency, and velocity.

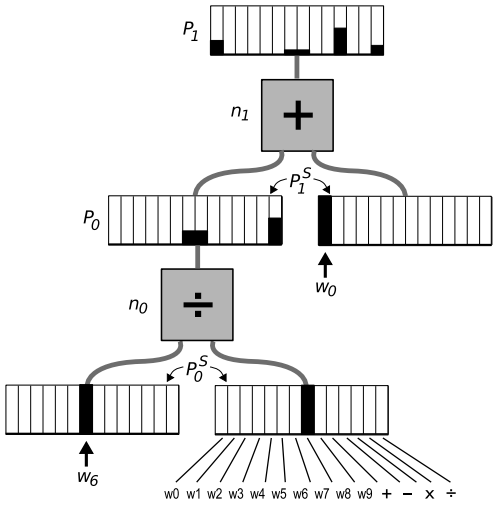
In recognition of this, we have used two different classifier architectures to capture patterns that occur over these two time-scales. To capture the local patterns of movement within a tap cycle, we used a variant of GP to evolve sliding window classifiers (see Section III-C). By representing patterns as mathematical (or more generally, programmatic) expressions, GP enables considerable flexibility in the way in which patterns are defined. For this reason, it has often been shown to outperform methods with more constrained representations [34]. It also leads to models that are relatively interpretable, a characteristic which is important for diagnostic classifiers, and which motivates the analysis later on in this paper. We have previously used this approach to induce classifiers for several biomedical diagnosis problems, including our earlier work on Parkinson's diagnosis [15], the discrimination of Raman spectra for cancer diagnosis [19], [35] and the classification of line drawings for the assessment of visuo-spatial ability [20].

GP is a useful technique for evolving static classifiers that describe static features. However, to capture the dynamical patterns of movement that occur over longer time-scales, we ideally want a classifier that is also dynamical. Whilst there have been successful attempts to introduce dynamical features to GP [36], in general features such as loops and memory remain fragile within an evolutionary context. As a consequence of this, there has been interest in using robust dynamical systems to represent computation within evolutionary systems. These computational dynamical systems (CDS) [37] include various kinds of recurrent neural network (RNN) and cellular automata, but also architectures motivated by the low-level biochemical networks that are directly exposed to evolution within biological systems. This includes our own work on artificial biochemical networks (ABNs) [38].

Several forms of CDS, including RNNs [39] and reservoir computers [40], have previously been applied to time series classification. In a preliminary study [41], we looked at whether ABNs can be used to separate neurological time series data, and found that they perform better at this task than a comparable RNN. Following from this, in this work we have taken a closer look at how dynamical ABN classifiers (described in Section III-D) can be used to recognise dynamical patterns occurring over a period of multiple taps, and how these complement the static GP classifiers that identify patterns occurring within single tap cycles.



(a) A 4×4 IRCGP grid representing the expression: $(w_2 + w_6)/(w_0 + w_6/w_7) + (w_2)(-w_7)(w_9)$



(b) Interconnections are the result of matching between functionality profiles. A node's functionality profile is derived from its own function and the functionality profiles of the sub-expressions bound to its inputs. For example, the functionality profile of the ÷ node is a weighted vector sum of its own function (the rightmost element in the functionality profile — see key at bottom) and the functionality profiles corresponding to the window offsets (w_6 and w_7) bound at its inputs.

Fig. 3. Examples of an Implicit context Representation CGP solution, showing (a) the arrangement of nodes in a Cartesian grid and (b) how the implicit context between two of the nodes is represented using functionality profiles.

B. Evolutionary Algorithm

We use the same evolutionary algorithm to evolve both the GP and the ABN classifiers: a standard generational EA with a population size of 200, a generation limit of 100, tournament selection (tournament size 4) and elitism (size 1), with child solutions generated using uniform crossover and mutation in the ratio 1:4. The objective in both cases is to accurately discriminate the acceleration time series recorded from PD patients from those of age-matched controls. We use the area under the ROC curve as a fitness function to measure this (see Section III-E for details), and use independent training, validation and test sets in order to obtain a reliable measure of classifier generality (see Section III-F).

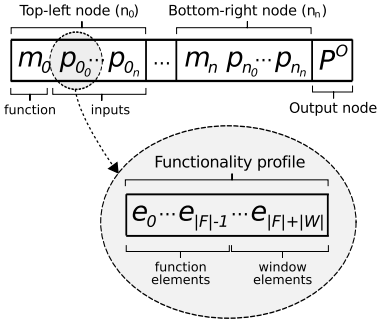


Fig. 4. Genetic encoding of an IRCGP expression.

C. Sliding Window IRCGP Classifier

IRCGP (Implicit context Representation Cartesian Genetic Programming) [42] is a graph-based GP system which uses the notion of implicit context [43]–[45] to provide positional independence to evolving solutions. IRCGP is a variant of Cartesian GP (CGP) [46]. Like CGP, an IRCGP solution consists of an n-dimensional grid (where n is typically 1 or 2) in which each grid location contains a function, and program inputs and outputs are delivered to and taken from specific grid locations (see Fig. 3a). However, unlike standard CGP, interconnections between functions, inputs and outputs are specified in terms of a component's *functionality profile*: a vector describing the component's functional context within the program (see Fig. 3b). Since functionality profiles are independent of grid position, this means that a program's behaviour is more likely to be preserved when variation operators modify a component's absolute or relative grid position. In particular, this has been shown to improve performance when crossover operators are used [45], [47]. In this respect, the goal of implicit context is similar to that of homologous [48] and semantic [49] crossover operators in GP; however, rather than changing the mechanics of how programs are recombined in order to preserve function, implicit context maintains function by reducing the impact of standard crossover operators.

Formally, an IRCGP expression is a tuple $\langle F, W, N, O\rangle$:

F is the function set $\{f_0, ..., f_n : \mathbb{R}^n \rightarrow \mathbb{R}\}$.
W is the input window $\{w_0, ..., w_n : \mathbb{R}\}$.
N is a set of nodes $\{n_0, ..., n_n : n_i = \langle m_i, P_i, S_i, P_i^S\rangle\}$, where:

- $m_i \in F$ is the node's function.
- P_i is the node's functionality profile.
- $S_i \subset N \cup W$ are the node's input sources.
- P_i^S are the node's input functionality profiles, such that $|P_i^S| = |S_i|$.

P^O is a functionality profile describing the network's output node.

Note that indices, n, are used as bound variables in each case.

Functionality profiles are used to express the connections between nodes in an IRCGP expression, and prior to execution are resolved to absolute grid positions (or, for terminal nodes, input window offsets) using a bottom-up development process. This process iterates through all the nodes in sequence, from the first to last row in the first column, and then similarly through the remaining columns. For each node, it then attempts to satisfy its input functionality profiles P_i^S by identifying downstream nodes S_i with the closest matching functionality profiles P_i, in terms of Euclidean distance. After all the nodes' input functionality profiles have been satisfied, and the corresponding inputs connected to the closest matching downstream nodes, the network's output node is determined by finding the node with the closest match to P^O.

A functionality profile p is a vector $\{e_0, ..., e_n : 0 \leqslant e_i \leqslant 1\}$ where $|p| = |F| + |W|$ and each e_i is an element corresponding to a particular function or window offset. A node's functionality profile P_i is defined recursively as the mean of its own function and the functionality profiles associated with its inputs:

$$P_i = 0.5P^{m_i} + 0.5\overline{P_i^S} \tag{1}$$

where P^{m_i} is a functionality profile in which the element corresponding to the node's function is set to 1 and all other elements are set to 0. Hence, P_i represents the relative, depth weighted, occurrence of functions and window offsets within the directed acyclic graph of which it is the head node, assuming its input functionality profiles are fully satisfied (i.e. matched to downstream nodes whose functionality profiles are exact matches) during the development process (see Fig. 3b). In practice, it is unlikely that each node's functionality profiles will be fully satisfied. Nevertheless, their inputs will be bound to the nodes (and corresponding sub-expressions) that most closely resemble the behavioural context declared by their input functionality profiles, and this will be the case even after they are recombined within child solutions, since the development process takes place before each new child solution is evaluated. This, in turn, promotes a process of gradual change, which has been shown to improve the performance of crossover [44] as well as leading to other beneficial evolutionary behaviours [45].

Following [20], IRCGP expressions use a function set F consisting of the standard arithmetic functions $\{+, -, \times, \div, \text{mean}, \min, \max, \text{mod}\}$. A subject's movement data is passed to the IRCGP classifier in the form of a real-valued time series of length l. This is then input to the evolved expression via a sliding window of length $|W|$. Hence, the evolved expression produces a real-valued output for each of $(l-|W|+1)$ overlapping time windows within the time series. The output of the classifier is the mean of these window values, reflecting the mean occurrence of the evolved pattern within the subject's movement data. Window sizes in the range of 10–20 are used, sufficient to cover a single tapping motion for an average subject.

1) Evolution: During evolution, IRCGP expressions are linearly encoded as shown in Fig. 4. The mutation rate is 6% for node functions, m_i, and 4% for the elements of functionality profiles, e_i. Real-valued elements (constants and the elements of functionality profiles) are mutated using a Gaussian distribution centred around the current value. A standard uniform crossover operator is used, with crossover points occurring with p=0.15.

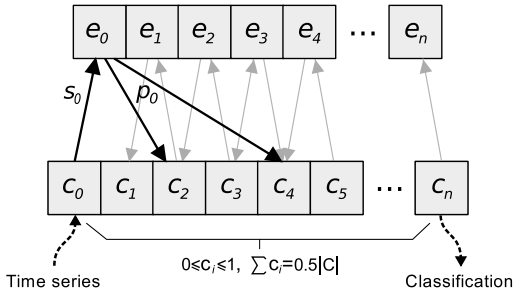


Fig. 5. An artificial metabolic network processing time series data. The time series is delivered one value at a time by setting $I_C = \{c_0\}$ and the final classification is read from $O_C = \{c_n\}$. The input signal is propagated both directly, via e_0, and indirectly, through the system's conservation law.

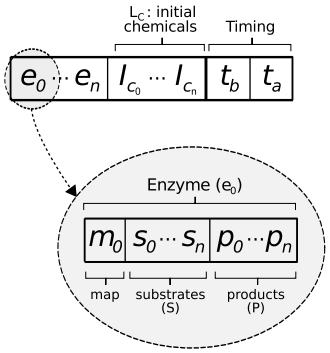


Fig. 6. Genetic encoding of an artificial metabolic network.

D. Artificial Biochemical Networks

Artificial biochemical networks (ABNs) are computational dynamical systems whose form and function are motivated by the biochemical networks found within biological cells. Like other computational dynamical systems (CDSs) [37], they display complex, time-varying behaviours that can be usefully applied to a range of computational tasks [50], [51]. ABNs have a number of architectural similarities with other connectionist CDS models, such as recurrent neural networks (RNNs) and cellular automata. However, they also have important differences, motivated by prominent patterns of organisation that occur in biochemical networks, rather than (for instance) neural networks. In previous work, we have looked at a number of variant ABN models, which include features such as dynamical nodal processes [51], self-modification [52], conservation laws [50], weak coupling between networks [53], and higher-order coupling [51]. We have found these architectures to be particularly useful for solving complex control problems, such as chaos control and legged robot locomotion [38], with different architectures being beneficial for different problems. For instance, self-modifying networks are useful when there is a requirement to switch dynamically between different behaviours. Notably, we have found the use of discrete maps within network nodes to be beneficial across a diverse range of problems [38], [41].

In our preliminary work [41], we found that a specific kind of ABN, the discrete map artificial metabolic network (AMN), achieved the highest classification accuracies when separating neurological time series data. An artificial metabolic network (AMN) is an abstract model of a cell's metabolism, capturing the idea of a set of enzyme-mediated reactions manipulating the concentrations of a set of mass-balanced chemicals over a period of time. In more familiar connectionist (i.e. neural network) terms, an AMN resembles an RNN in which activation levels are shared between neurones, and the sum of activation levels is maintained at a fixed level. In a discrete map AMN, enzyme-mediated reactions are modelled as non-linear iterative maps. These capture the dynamical complexity of the kinds of processes that occur in biochemical networks [54], but in a computationally efficient form. In our previous work, we have found them to be effective for exploring diverse dynamical behaviours, and thereby promoting behavioural diversity within evolving populations [38]. In our initial investigation of AMNs evolved for time series classification [41], we analysed the behaviour of a single evolved discrete map AMN, and found that it was highly sensitive to relatively small changes in its input: a sensitivity which appeared to result from the interplay between the chaotic behaviour of the discrete maps and the dampening behaviour of the conservation law.

Formally, an AMN is a tuple $\langle C, E, L_C, I_C, O_C \rangle$:

C is the set of chemical concentrations $\{c_0, ..., c_n : \mathbb{R}\}$.
E is the set of enzymes $\{e_0, ..., e_n : e_i = \langle S_i, P_i, m_i \rangle\}$, where:
- $S_i \subseteq C$ is the enzyme's substrates.
- $P_i \subseteq C$ is the enzyme's products.
- $m_i : S_i \rightarrow P_i$ is the substrate-product mapping.

L_C is an indexed set of initial chemical concentrations, where $|L_C| = |C|$.
$I_C \subset C$ is the set of chemicals used as external inputs.
$O_C \subset C$ is the set of chemicals used as external outputs.

AMNs are executed as follows: First, their chemical concentrations are initialised from L_C. During the course of execution, external inputs are delivered by explicitly setting the concentrations of chemicals indicated in I_C at appropriate intervals. At each time step, the enzymes synchronously modify the chemical concentrations according to their defined mappings. To maintain mass balance, the chemical concentrations are then uniformly scaled so that they sum to $0.5|C|$. Chemicals which have reached saturation ($c = 1$) and those which are not present in the chemistry ($c = 0$) remain unchanged, preserving these special states. At the end of execution, outputs are captured from the final concentrations of the chemicals specified in O_C.

An acceleration time series is input to an AMN by setting the concentration of the first chemical (c_0). The time series is delivered to a network one value at a time, each followed by t_b iterations of the network. Once the whole time series has been delivered, the network is executed for another t_a iterations in order to allow the dynamics to settle. At this point a single output value is read from the final concentration of the last chemical (c_n). Using a suitable threshold, this output value can then be interpreted as the network's classification for the time series. The settling parameters, t_b and t_a, are both evolved with the network.

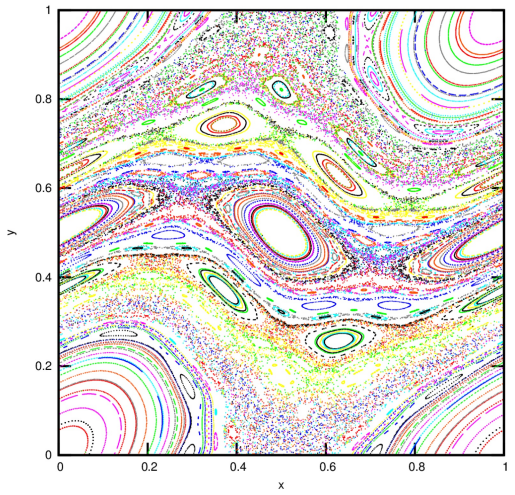


Fig. 7. Chirikov's standard map, which maps points within the unit square and is a general model of conservative discrete dynamical systems with co-existing ordered and chaotic dynamics. For $k = 1.1$ (plotted), ordered dynamics are mostly found in the centre and the corners, and chaotic dynamics mostly occur in the lower and upper regions, excluding the corners. Sampled trajectories for 200 initial points are shown.

1) Discrete maps: Following [41], we use the following set of behaviourally-diverse discrete maps to implement the enzymes' substrate-product mappings: the logistic map, Chirikov's standard map, the baker's map, and Arnold's cat map. Between them, these maps are capable of expressing a wide range of dynamical phenomena, many of which are known to occur within natural systems.

The logistic map [55] is a model of biological population growth which displays either ordered or chaotic behaviour depending upon the value of a parameter $r \in [0, 4]$:

$$x_{n+1} = r x_n (1 - x_n) \tag{2}$$

In order to allow the map to switch between dynamical regimes during execution, we use a tunable variant of this map in which r is set using an extra input, rather than remaining constant.

Chirikov's map [56] is a model of Hamiltonian systems whose phase spaces have co-existing ordered and chaotic regimes (see Fig. 7). The dynamics move from majority-ordered to majority-chaotic as the parameter $k \in [0, 10]$ increases:

$$\begin{aligned} x_{n+1} &= (x_n + y_{n+1}) \bmod 1 \\ y_{n+1} &= (y_n - \frac{k}{2\pi} \sin(2\pi x_n)) \bmod 1 \end{aligned} \tag{3}$$

The baker's map [57] and Arnold's cat map [58] are both archetypal models of chaotic phenomenon that occur in a range of systems. They are defined, respectively:

$$(x_{n+1}, y_{n+1}) = \begin{cases} (2x_n, y_n/2) & 0 \le x_n \le \frac{1}{2} \\ (2 - 2x_n, 1 - y_n/2) & \frac{1}{2} \le x_n < 1 \end{cases} \tag{4}$$

$$(x_{n+1}, y_{n+1}) = ([2x_n + y_n] \bmod 1, [x_n + y_n] \bmod 1) \tag{5}$$

2) Evolution: During evolution, AMNs are linearly encoded as shown in Fig. 6. Each substrate-product mapping, m_i, encodes both the choice of discrete map and any associated parameters. All components of the AMN are subject to point mutation, at a rate of 6% per component. The number of chemicals is fixed at 10, and the number of enzymes has a lower bound of 1, with no upper bound. Crossover points (p=0.15) always fall between enzymes. Real-valued elements (L_C and parameters associated with non-tunable discrete maps, i.e. r for the logistic map and k for Chirikov's map) are mutated using a Gaussian distribution centred around the current value. Enzymes and enzyme substrates may be added or removed (p=0.015/element), the former either randomly (p=0.5) or by duplicating an existing enzyme or substrate (p=0.5).

E. Classifier Evaluation

A classifier's ability to correctly assign class membership to previously unseen data is known as its *predictive power*. There are many ways of measuring predictive power [59]. Underlying many of these is the notion of a *confusion matrix* (or *contingency table*), a table recording the number of data points correctly and incorrectly mapped to each class. In the binary case, where data points can be considered to be positive and negative examples of one of the two classes, the confusion matrix is a two-by-two table showing the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) predictions.

Below we review the metrics used in this paper.

1) Accuracy: The simplest, and most obvious, measure of predictive power is the proportion of the data set which is correctly classified, i.e. the proportion of input cases which are correctly mapped to their respective class labels. This metric is known as *accuracy*. For a binary classifier, it is defined:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

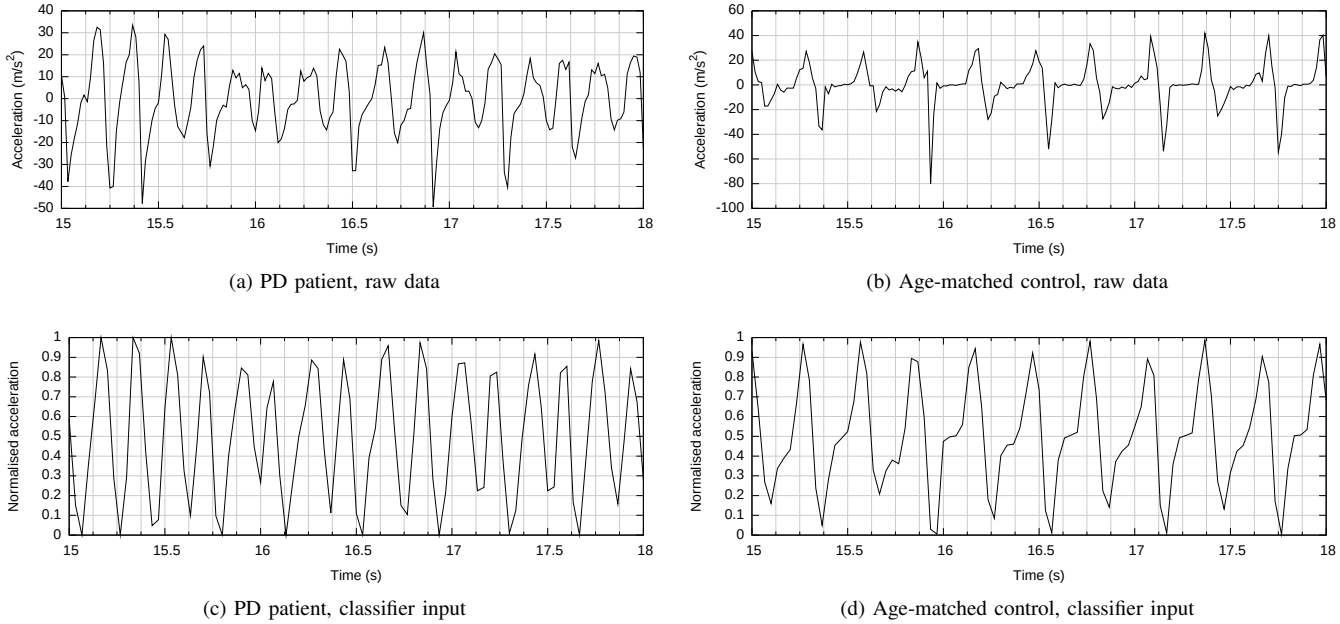
Accuracy is commonly used to train classifiers, and can be readily applied to both binary and multi-class classifiers. However, it is sensitive to class size distribution, and is generally a poor choice when there is significant class size variation [60].

2) Specificity and sensitivity: Two class size insensitive metrics derived from a binary confusion matrix are specificity and sensitivity. *Specificity*, also known as the *true negative rate* (TNR), is the probability that a negative classification will be given for a negative data point. *Sensitivity*, the *true positive rate* (TPR), is the probability that a positive classification will be given for a positive data point. They are calculated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

Sensitivity and specificity capture two, often opposing, aspects of predictive power: the ability to recognise all positive examples and the ability to reject all negative examples. The relative importance of these two activities depends upon the context in which the classifier is being used. For example, during medical screening, it is important to have high sensitivity, so that no cases are missed. At a later stage, when the presence



(a) PD patient, raw data

(b) Age-matched control, raw data

(c) PD patient, classifier input

(d) Age-matched control, classifier input

Fig. 8. Examples of a PD patient and an age-matched control performing finger tapping over an interval of 3 seconds, showing (a–b) raw acceleration data, and (c–d) corresponding acceleration sequences after preprocessing.

of a disease is being confirmed, it is important to have high specificity, so that patients without the disease do not undergo unnecessary treatment.

3) ROC metrics: Many types of classifier (including those used in this study) produce continuous-valued outputs, which must then be mapped to class labels using thresholds. Consequently, different trade-offs between specificity and sensitivity can be achieved by varying these thresholds. This is captured by a Receiver Operating Characteristic (ROC) curve, a plot of true positive rate (sensitivity) versus false positive rate ($1-$specificity) for all possible thresholds on the classifier's output range.

A number of single-valued summary statistics can be calculated from a ROC curve. The most common of these is the Area Under Curve (AUC), which is calculated by integrating the area under the ROC curve, typically using the trapezoid rule. AUC is equivalent to the probability that a positive data point will be given a higher output value than a negative data point [61]. As such, an AUC of 0.5 is equivalent to random classification. AUC is symmetrical, meaning that a classifier with an AUC of 1 has the same predictive power as one with an AUC of 0 (although with an inverted ordering of classes in its output range). Its relationship to probability means that the AUC is easy to interpret, making it a popular metric in medicine [62].

F. Training and Test Data Sets

Two-thirds of the clinical recordings are placed in a training set, which is used for fitness evaluation. The other third of the data (referred to collectively as the non-training data) is used to measure classifier generality. In situations where we wish to select the best-performing classifier from a series of evolutionary runs, the non-training data is further divided, approximately equally, into validation and test sets. In this situation, the validation set is used to identify the best-performing classifier, and the test set is used to give an unbiased measure of its discriminative power. See Table I for a summary of data set sizes. To compensate for any clinical deviations over the course of the trial, subjects are evenly distributed between data sets with respect to date of recording. To prevent overestimation of performance, left and right hand recordings are always kept together when used to train bilateral classifiers.

TABLE I
DATA SET SIZES

	Training	Non-training	Validation	Test
Patients	33	16	8	8
Controls	28	13	7	6

G. Preprocessing

Each sensor's translational (x, y, z) and rotational (elevation, azimuth, roll) coordinate data were collected every 1/60th of a second. For each time index, the Euclidean distance between the two sensors was calculated, generating a sequence of sensor separations over time for each subject. These were then converted into acceleration time series. An initial investigation suggested that classifiers trained on raw acceleration data were sensitive to signal noise, and would also converge suboptimally to simple classifiers based on amplitude alone, a variable which is moderately predictive of PD (see Table II). To mitigate this, the data was preprocessed prior to classifier training. First, to remove noise, the data was down-sampled by a factor of two and a moving average filter of size 2 was applied. The acceleration data was then truncated to one standard deviation around the mean and scaled uniformly

to the interval $[0, 1]$ to remove information about absolute amplitude. Examples of raw and preprocessed acceleration time series are shown in Fig. 8.

IV. Results

A. Baseline Measures

Previous studies in the medical literature have also considered using recordings of finger tapping as a basis for diagnosing movement disorders. These have generally focused upon gross features of movement data, such as mean amplitude and velocity. In [63], for example, the authors note a fairly strong correlation ($\sim$0.8) between UPDRS score and both velocity and spectral power within gyroscopic recordings of finger tapping from 40 patients and 14 controls. Similarly, using various gross features, including amplitude fatiguing, Ling and colleagues were able to identify differences in finger tapping performance between patients with PD and patients with progressive supranuclear palsy not captured by UPDRS finger tapping scores [64].

TABLE II
Baseline AUCs for gross data features

Variable	Dominant hand	Non-dominant hand
Mean amplitude	0.78	0.73
Mean speed	0.73	0.67
Amplitude fatiguing ratio	0.58	0.60
Tapping frequency	0.50	0.58

1) Gross Features: To test the utility of gross features as a basis for classification, we computed tapping frequency, mean amplitude, mean speed, and amplitude fatiguing ratio (the ratio of amplitudes for the first 5 taps and the last 5 taps) for each subject, and then calculated the AUCs when each of these measures was used to separate patients from controls. The results are shown in Table II. Mean amplitude offers the best discrimination between patients and controls, with an AUC of 0.78 on the dominant hand. Whilst this does suggest a significant correlation between tapping amplitude and PD, the level of discrimination is too low to be useful for diagnosis. Speed appears to be less useful as a predictor of PD, and the other metrics—frequency and fatiguing—are relatively ineffective as a basis for classification, and it is unlikely that a significantly better classifier could be constructed through linear combinations of these features.

2) Parkinsonian Tremor: We also looked at the incidence of rest tremor, since this is a common symptom of PD and can be readily measured using spectral analysis methods. To verify the occurrence of tremor within the 4-6Hz frequency range indicative of PD [21], we carried out a Fourier analysis of the rotational components of the data, in which tremor is most often seen. Power spectrum and confidence limits were computed using the method described in [65]. These indicated significant peaks in the 4-6Hz range in one or both hands for 15 (31%) patients (see example in Fig. 9). Of these, 5 displayed tremor in their dominant hand, 9 displayed tremor in their non-dominant hand, and 1 patient exhibited tremor in both hands. This supports current understanding that the limited incidence of rest tremor means that it can not by itself be used to diagnose PD [66].

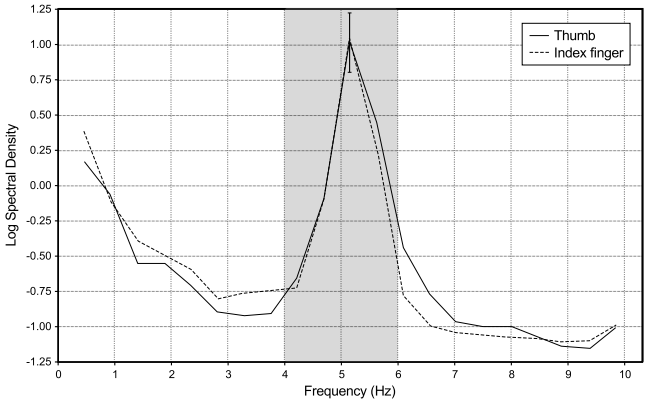


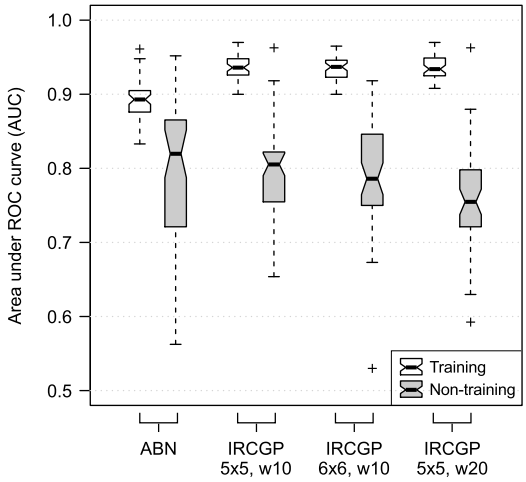
Fig. 9. Power spectral density of a patient's hand movements at rest, showing a clear peak in the 4-6Hz band, indicative of Parkinson's.

B. Evolved Classifiers

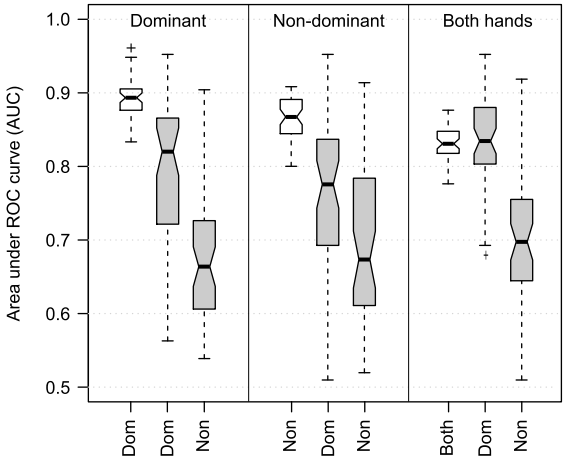
Table II shows that dominance has a significant effect upon discriminatory power, with gross features of the data having lower predictive accuracy for the non-dominant hand. Consequently, we limited our initial investigation to classifiers trained on preprocessed finger tapping data collected from each subject's dominant hand. Fig. 10a shows the distribution of AUC scores for the best evolved classifiers from each of 50 runs.

For both classifier architectures, a number of classifiers were evolved with training and non-training AUCs of 0.9 and greater. This suggests that both architectures are able to express patterns that discriminate well between PD patients and controls. Overall, the ABN runs produced more high-AUC classifiers than the GP runs, although the best-performing classifier was a GP expression, with an AUC of 0.96 on the non-training data. Fig. 11 shows the ROC curves for the ABN and GP classifiers with highest discrimination on the validation set. In both cases, the test set AUC is very high, indicating that these classifiers generalise well to unseen data. Fig. 10a also shows that, in general, the size of the IRCGP grid and the length of the matching window has a relatively small effect upon the ability of the evolutionary algorithm to find classifiers.

Although most runs led to classifiers with high training AUCs, the wide non-training set distributions shown in Fig. 10a show that a number of evolved classifiers had poor generality. For the GP classifiers, this was caused by over-learning in a number of runs, with the non-training set AUCs peaking early whilst the training set AUCs continued to increase. This also explains why the training set AUCs are higher for the GP runs than for the ABN runs. For the ABN classifiers, poor generality was caused by an evolutionary trend towards parsimonious solutions. Below a certain size (about 3 discrete maps), we found that solutions displayed high fitness but poor generality (see [41] for a more in-depth discussion of this phenomenon). In each case, early stopping and solution

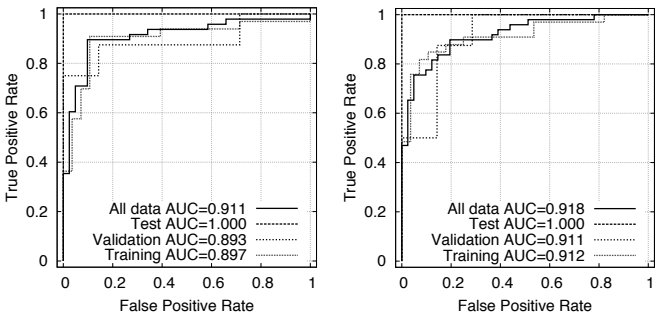


(a) Comparison of ABN and IRCGP classifiers, also showing effect of grid and window (w) size for IRCGP classifiers.



(b) Comparison of ABN classifiers evolved on training data from (left to right) **Dom**inant, **Non**-dominant and **Both** hands. In each case, the first boxplot shows the results from training and the other two show the diagnostic power for the dominant and non-dominant hand recordings in the non-training data set.

Fig. 10. Diagnostic power of evolved classifiers on both the training (white) and non-training (grey) sets. Notched box plots show summary statistics over 50 runs. Overlapping notches indicate when median values (thick horizontal bars) are not significantly different at the 95% confidence level.



(a) ABN Classifier (b) GP Classifier

Fig. 11. ROC curves for the most discriminative ABN and GP classifiers.

size limits failed to improve generality, suggesting that these

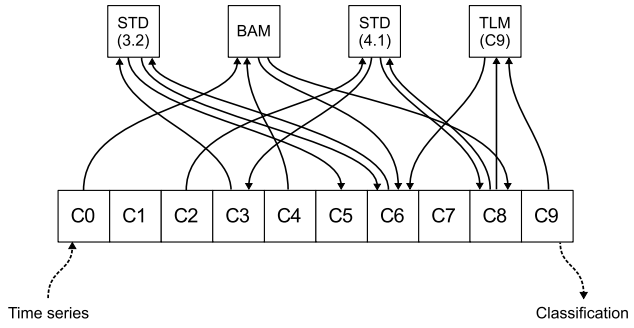


Fig. 12. Evolved ABN Classifier, comprising two standard maps (STD), one baker's map (BAM), one logistic map (TLM), and 10 chemicals.

behaviours are due to deceptive fitness landscapes.

C. Handedness

PD is typified by asymmetric onset, and a number of recent studies have suggested that there is a positive correlation between a patient's handedness and the side of their body on which symptoms first present [67], [68]. Our investigation of gross metrics in Table II tends to support this, showing that these provide slightly higher discrimination for a subject's dominant hand.

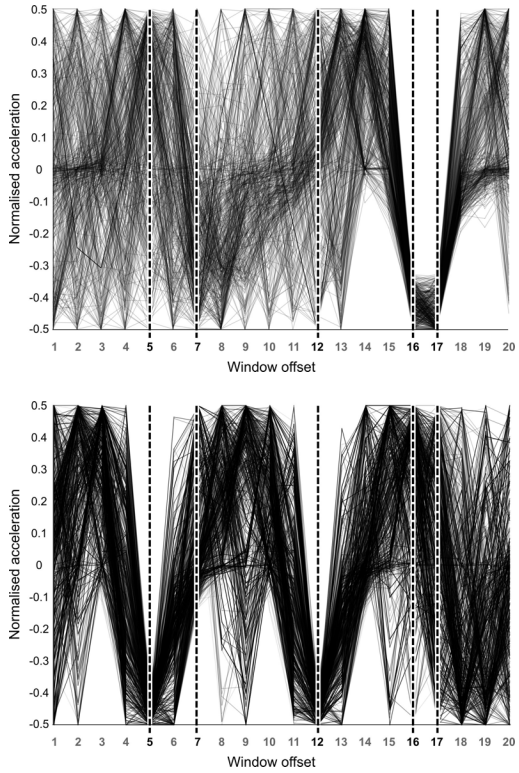
To test this relationship further, we separately trained classifiers using data from the dominant, non-dominant and both hands. Fig. 10b shows that the evolutionary algorithm found it significantly easier to find high fitness classifiers for the dominant hand. This pattern is even more pronounced for discrimination on the data not used for training, suggesting that it is much harder to perform diagnosis when using non-dominant hand data, and providing strong support for the findings of [67].

Interestingly, classifiers trained on the non-dominant hand still generalise well to the dominant side. This suggests that the same pattern is found on both sides, but with greater incidence or fidelity on the dominant side. It is also notable that the distribution of classifiers trained on both hands, then re-evaluated on the dominant hand in the non-training data, shows less indication of over-learning than those trained solely on the dominant hand. We can speculate that, by making the pattern harder to find, this reduces early convergence of the population.

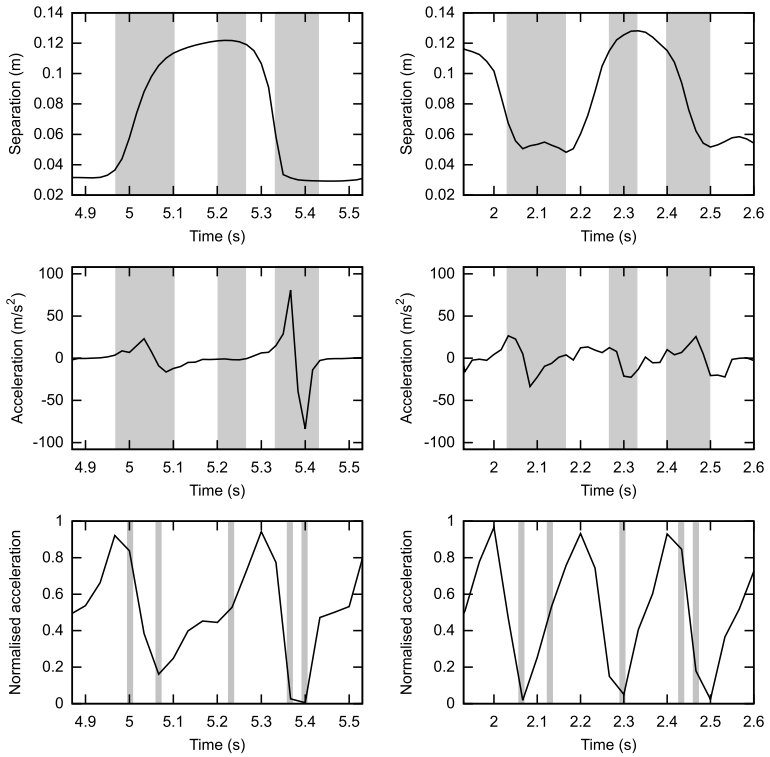
D. Behavioural Analysis

Fig. 12 shows an evolved ABN, showing how it is relatively simple in terms of description length, comprising 4 discrete maps and 10 chemicals. However, due to its dynamical nature, and the non-trivial collective behaviour that results from coupled discrete maps [69], it is extremely difficult to infer its functional behaviour from its static description. Eq. 9 gives an example of an evolved expression used by a GP classifier, where w_i is the value at offset i within the sliding window and sub is a sub-expression reused by the CGP graph:

$$\begin{aligned} \mathrm{out} &= (0.531 + \mathrm{sub})\mathrm{sub} \\ \mathrm{sub} &= \frac{\max\{w_{12}, w_5\}}{max\{w_{16}, w_{17}\} + 0.078w_7} \end{aligned} \tag{9}$$



(a) Overlay of data windows which the evolved expression classifies as (top) normal and (bottom) abnormal. The window offsets used as inputs to the evolved expression are shown by broken vertical lines.



(b) Examples of individual data windows classified as (left) highly normal and (right) highly abnormal. The corresponding patterns in the (top) raw separation and (middle) raw acceleration data are also shown, with grey regions indicating the parts of the window that contribute to the active classifier inputs.

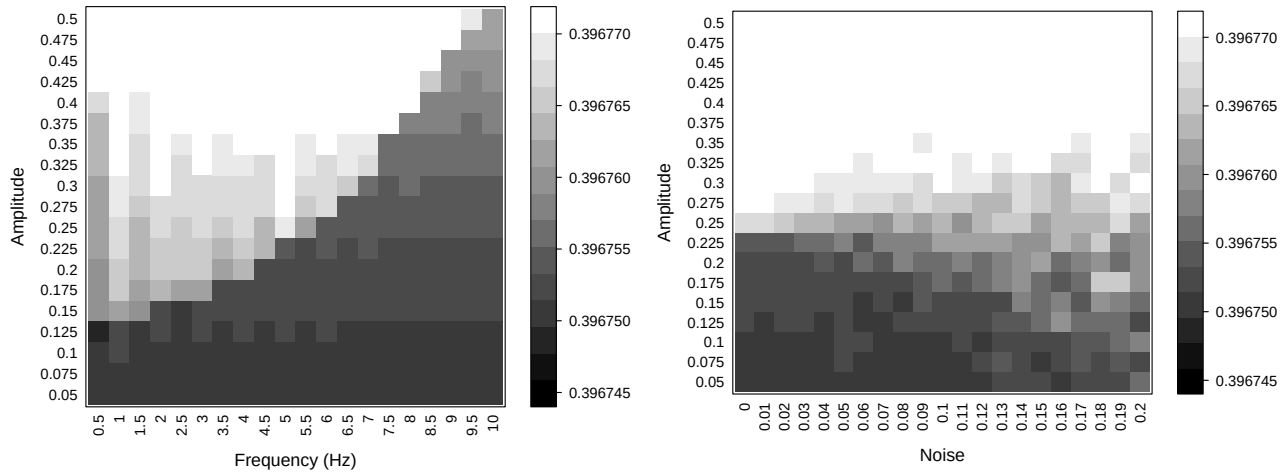
Fig. 13. Analysis of an evolved GP classifier, $\mathsf{AUC}_{\mathsf{all}} = 0.919$, optimal output threshold $\sim$3.5.

This is also surprisingly simple, but nevertheless it is still difficult to understand the pattern of movement it is describing.

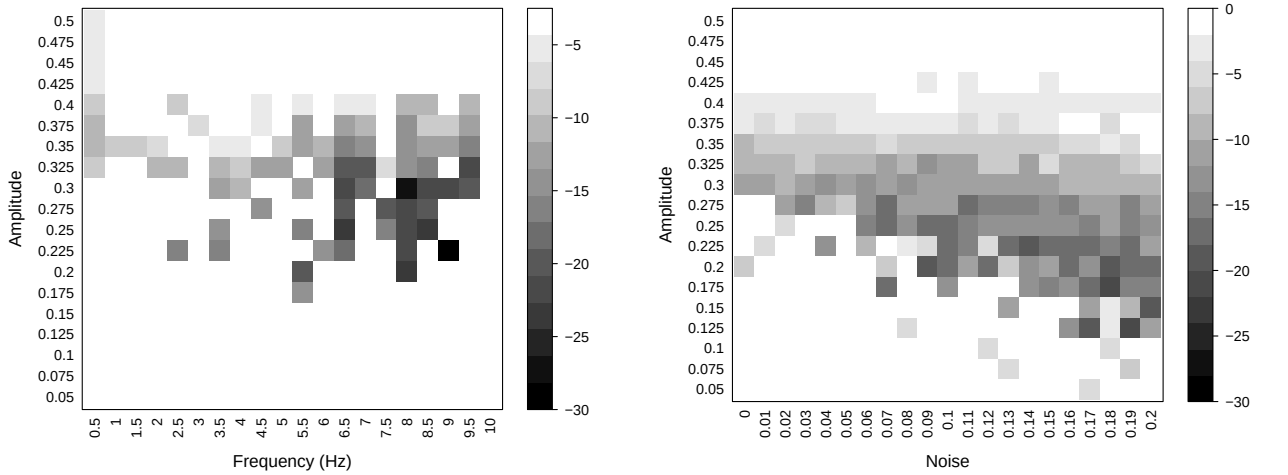
1) Local Patterns: For the GP expression, we can gain insight into its behaviour by looking at the time series windows which receive either a low or a high output from the expression, since these correspond to the local patterns of acceleration which are identified as abnormal or normal in the PD and control sequences, respectively (or vice versa, depending upon the ordering of classes in the classifier's output range). Fig. 13a shows an overlay of all the time series windows that are classified as particularly normal or particularly abnormal by the GP expression in Eq. 9. Although there is a degree of fuzziness, it can be seen that there is a distinct over-represented pattern in each case: a sinuous pattern of acceleration and deceleration for normal matches, and a pattern centred around two closely-spaced deceleration peaks for abnormal matches. To clarify the meaning of these patterns, Fig. 13b shows examples of two windows that are classified as highly normal and highly abnormal. It can be seen that the sinuous pattern noted in Fig. 13a appears to correspond to a smoothly-changing opening and closing movement. In addition, the closing deceleration is significantly larger than the opening deceleration, which reflects the inelastic collision as the two fingers collide at the end of the movement. The double deceleration pattern in the abnormal match, by comparison, corresponds to a jerky pattern of motion in which the final deceleration is of a similar magnitude to the opening deceleration. Notably, this jerky motion resembles one of the known symptoms of PD, cog-wheel rigidity. However, perhaps more interesting is the abnormal relationship between opening and closing deceleration, which we also observed in other windows labelled as highly abnormal. This indicates that PD patients are slowing their fingers prior to the inelastic collision, which in turn is indicative of a breakdown in sensory feedback, a feature of PD which has only recently received significant interest in the medical literature [70].

2) Global Patterns: We cannot perform this kind of analysis for ABNs, since they operate at a sequence-level rather than a window-level. Instead, we can characterise an ABN's transfer function by measuring its response to synthetic time series data with known properties—particularly properties that are expected to have diagnostic relevance for PD, such as amplitude, frequency and irregularity. Analysing highly-discriminative ABN classifiers in this way shows that they have diverse dynamical responses, suggesting that they recognise a range of different patterns when carrying out diagnosis. Fig. 14 gives examples of responses for 3 different highly-discriminative ABNs when perturbed with sine waves of differing amplitude, frequency and levels of added noise (which approximate the jittery movements of some PD patients).

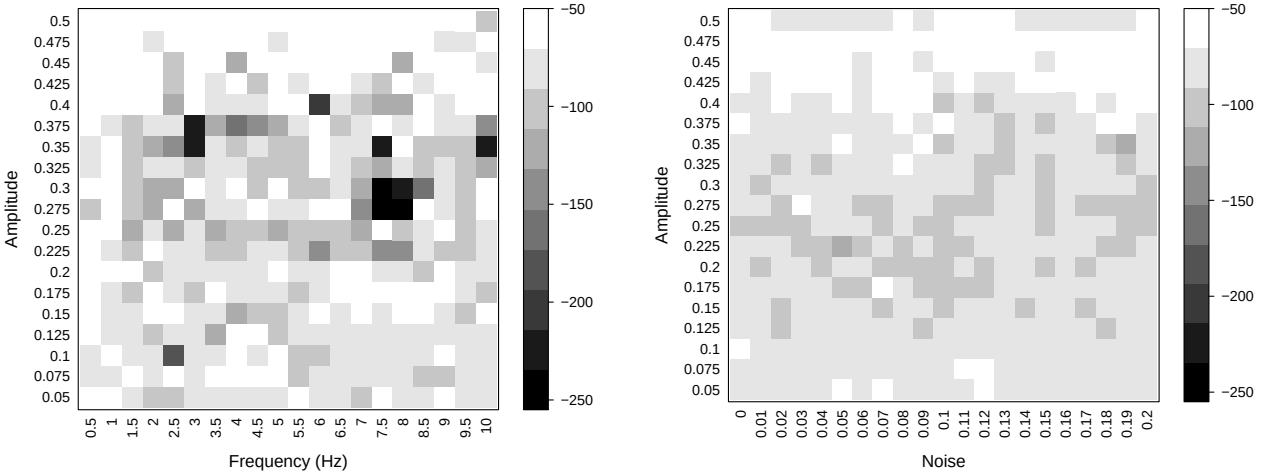
The ABN in Fig. 14a has a relatively clear amplitude-frequency response. In terms of distance from the decision



(a) ABN1: $\mathsf{AUC}_{\mathsf{all}} = 0.893$, optimal output threshold $\sim$0.396770



(b) ABN2: $\mathsf{AUC}_{\mathsf{all}} = 0.901$, optimal output threshold $\sim 1 \times 10^{-4}$.



(c) ABN3: $\mathsf{AUC}_{\mathsf{all}} = 0.911$, optimal output threshold $\sim 1 \times 10^{-68}$.

Fig. 14. Examples showing how different evolved ABN classifiers respond to sinusoidal signals of differing amplitude, frequency and noisiness. Note that the output scales for (b) and (c) are logarithmic, with values shown $\log_{10}$. Output responses that fall within the classifiers' non-PD range are shown in white; those in the PD range are shown in grey, with intensity proportional to the magnitude of the classifier's response.

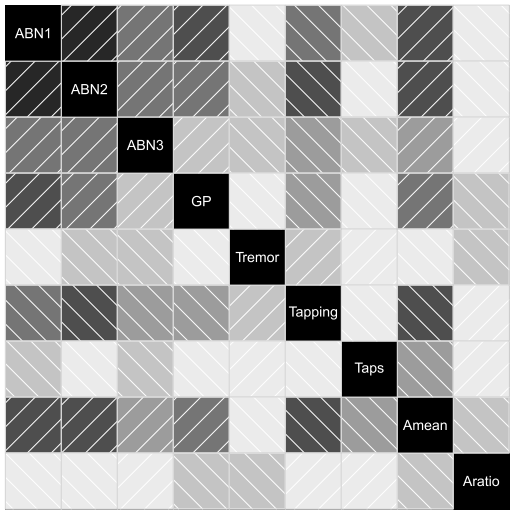


Fig. 15. Correlogram showing Spearman correlation coefficients between ABN and GP classifiers, UPDRS tremor and tapping scores, tapping frequency ('Taps'), mean amplitude ('Amean'), and the ratio of mean amplitudes during the first 5 taps and the last 5 taps ('Aratio'). Shading intensity shows the magnitude of correlations. Line direction indicates the direction of correlations (bottom-left to top-right = positive; top-left to bottom-right = negative).

boundary, it responds most strongly to low amplitudes at low frequencies and all amplitudes at high frequencies. In the preprocessed movement data, regions of low amplitude tend to correspond to fatigue and irregular tapping, which are both seen in PD patients. The high frequency part of the plot is well outside a subject's normal tapping frequency, and the strong response in this region may reflect the presence of multiple acceleration peaks during a single tap. Again, this correlates well with the cog-wheel like motion seen in PD patients. The presence of noise in the synthetic sine waves also has a slight effect on the ABN's response, increasing the amplitude at which signals are classified as PD.

The ABN in Fig. 14b has a less clear response to amplitude and frequency, classifying sine waves as PD based on the incidence of intermediate amplitudes, particularly at high frequencies. Again, this may indicate a response to cog-wheel like motion. The response to noise is more clear, with the ABN responding to a wider range of amplitudes as noise increases. This would suggest that the ABN responds to irregular or jittery behaviour indicative of poor motor control, even when the amplitude of tapping appears relatively normal.

By comparison, Fig. 14c shows no clear pattern in its response to either frequency or noise, although it does tend to classify high amplitude signals as non-PD, which reflects consistent tapping. This is the most-discriminative of the three solutions, and presumably its discriminatory ability is based upon other factors, such as waveform shape.

3) Correlations: This pattern analysis suggests that evolved classifiers respond to a variety of different signals within the tapping time series data. This is corroborated by Fig. 15, which shows the correlations between the output responses of the ABN and GP classifiers, various gross features of the finger tapping data, and also the clinician's UPDRS tremor and tapping scores.

ABNs 1 and 2 have well-correlated outputs. However,

TABLE III
PERFORMANCE OF SCALED MEAN ENSEMBLES

Ensemble	AUC_{all}
ABN1, ABN2, ABN3	0.922
ABN1, ABN2, ABN3, GP	0.949
ABN1, GP	0.936
ABN2, GP	0.933
ABN3, GP	0.959

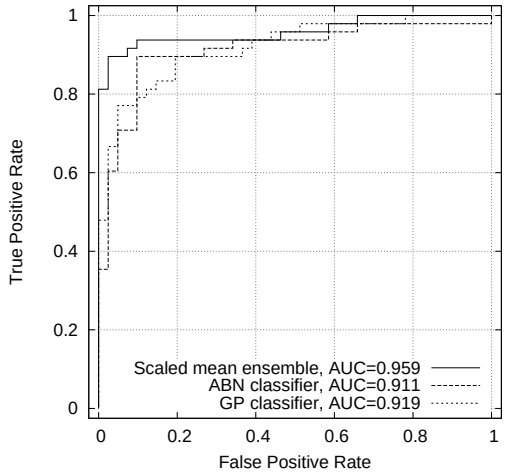


Fig. 16. ROC curves for an ensemble and its component classifiers.

they differ in their correlations with gross features and UPDRS scores, suggesting they reach similar classifications but through slightly different means. The GP classifier's outputs correlate fairly well with ABN1, reflecting our observation that both have strong responses to certain frequency components. By comparison, the outputs of ABN3 and the GP classifier have poor correlation, reflecting ABN3's minimal response to frequency components. Only the GP classifier has a significant correlation with the amplitude fatigue ratio, despite having no global view of the tapping data. The outputs of ABNs 2 and 3 both show small correlations with the UPDRS tremor score, suggesting that they may be responding to Parkinsonian tremor—however, the size of the correlation reflects the poor discriminatory power of tremor as an indicator for diagnosis. The outputs of all the classifiers show moderate to high correlations with the UPDRS tapping score, suggesting that the evolved classifiers are performing a broadly similar task to the trained clinicians.

E. Ensemble Classifiers

Since the classifiers appear to respond to different patterns in the finger tapping data, it seems likely that higher discrimination accuracy could be achieved by forming classifier ensembles. There are numerous ways of combining the outputs of classifiers. For continuous-valued classifiers such as these, one approach is to use the mean of the classifiers' outputs. This has the advantage of maintaining a continuous-valued output, and therefore allowing different thresholds to be chosen depending upon how the classifier is used. However, the output ranges

first need to be normalised. We do this by uniformly scaling the output range of each component classifier to the interval $[0, 1]$, i.e. the lowest output is mapped to 0 and the highest output is mapped to 1. For classifiers with a logarithmically-distributed output range, we first take the common logarithm of the output value.

Table III shows the performance of ensembles constructed from the classifiers analysed in the previous section. Whilst there is a small advantage to combining the different ABNs, the largest benefit comes from combining the two different types of classifiers, i.e. an ABN and a GP classifier. In particular, the best performance is achieved by combining the two classifiers which show the least correlation in Fig. 15. This reflects previous understanding that, when selecting classifiers to form an ensemble, best results can be obtained by selecting classifiers whose outputs are least correlated [71]. Fig. 16 shows how the resulting ROC curve improves upon the ROC curves of the component classifiers, particularly at the low false positive side. This, in turn, shows the benefit of using two different classifier models, one which is sensitive to relatively well-characterised local features, and one which is able to recognise less evident global features.

V. Discussion

A. Sample sizes

A common issue with medical studies is the use of small samples of the diseased and healthy populations [72]. Small samples are particularly commonplace in studies, such as this one, which require time-consuming and one-off clinical measurements. Nevertheless, our sample of 49 PD patients and 41 controls compares favourably with other recent PD studies, including Tsanas et al.'s use of SVNs to discriminate vocal recordings (32 PD, 10 controls) [18], Kim et al.'s study of discriminatory features during finger tapping (40 PD, 14 controls) [63], and Ling et al.'s study of discriminating movement characteristics (15 PD, 16 controls, 9 progressive supranuclear palsy) [64].

Small samples significantly increase the difficulty of classification problems. The smaller the sample size, the greater is the likelihood that an induced classifier will respond to a spurious pattern that is over-represented in the sample, rather than (the aim in this case) a biologically meaningful pattern. This can be seen in the wide non-training set distribution in Fig. 10a, an indication that a number of evolved classifiers do not generalise to previously unseen data. It also demonstrates a benefit of using evolutionary algorithms: the ability to explore a wide range of solutions, both within and between runs, increasing the chance of finding those with generality. However, this also underscores the importance of using separate validation and test sets, in order to correctly differentiate classifiers that generalise from those that do not.

In our case, we only used the test set once for each classifier architecture, in order to measure the generality of the two classifiers which performed best on the validation set. The resulting high level of discrimination (Fig. 11) is a strong indication that these classifiers have good generality with regard to the wider PD and neurologically healthy populations. Our behavioural analysis also supports this conclusion, showing that the classifiers are responding to features that we would expect to be diagnostically significant. Hence, the analysis of evolved classifiers provides a sanity check in addition to having the potential to inform wider clinical practice.

B. Clinical Perspectives

Fig. 16 suggests an accuracy of around 90-95%, depending on the chosen trade-off between sensitivity and specificity. This is similar to the accuracies achieved by expert clinicians when making a diagnosis, and considerably higher than those achieved in primary and community care [21]. However, it should be noted that this is not the same problem faced by clinicians when making an initial diagnosis. In one sense it is easier, since clinicians must discriminate between a number of related neurological conditions when reaching a diagnosis. On the other hand, it is harder because the classifier has access to less information and, because the patients are on medication, their symptoms are harder to discriminate.

The accuracy of the evolved classifiers in detecting the motor symptoms of PD suggests that our approach could be used to develop new diagnostic and monitoring tools. These tools would offer several important advantages to the clinician and clinical researcher. Data could be collected efficiently while the patient performs the standard clinical tests of bradykinesia and rest tremor that are already a widely-accepted part of PD assessments. Unlike clinician judgement of performance on these clinical tests, distinct performance factors could be derived. These factors may prove useful for differentiating PD from other diseases (e.g., progressive supranuclear palsy), for characterizing the motor subtypes of PD, and for measuring an individual patient's unique motor symptom profile. In addition, symptoms could be characterised on a fine scale, which may be more sensitive to small but important changes in motor features not captured by traditional scales. While this advantage has clear importance for clinical care, it could also be relevant to clinical trials, because outcome measures that are sensitive to small treatment effects do not require as large sample sizes to establish treatment efficacy. Lastly, an advantage of this method is that scores are objectively derived and thus not impacted by subjective clinician judgment.

Although the present results are promising, more research is needed before these classifiers are ready to be applied as a clinical tool. In particular, research on newly diagnosed or prodromal cases, as well as on patients with other diseases that are often confused with PD, will be needed to determine how useful the classifiers are for early and accurate diagnosis. Studies comparing patients when on and off their dopaminergic therapies will be necessary to determine the sensitivity of the classifiers to treatment response. Inter-rater reliability should also be evaluated. All of these studies will need to contrast the efficacy of the classifiers in comparison to traditional assessments, like the UPDRS, to determine whether the classifiers provide added utility.

C. Biomedical Data Mining

A distinctive aspect of this work is the use of dynamical classifiers, i.e. artificial biochemical networks. Biomedical

signals, such as those produced by the human nervous system, are complex, non-linear and non-stationary—in a word, dynamical. Despite this, most work on classification of biomedical time series data is currently done using static classifiers such as support vector machines [73] and feed-forward neural networks [74] operating on data windows or other extracted static features. There has been some work on applying static classifiers to dynamical features, such as those produced by spectral analysis [75] and time-delay embedding [76]. However, in many cases there is little *a priori* understanding of the kind of dynamical features we are looking for—and, in these cases, it arguably makes more sense to fit a general dynamical model to the data, such as a recurrent neural network [39], a reservoir computer [40], or in our case, an ABN. We can see the advantage of this approach in our work, where the ABN classifiers are considerably more diverse and, on average, more discriminative than the static GP classifiers.

A more general advantage to using evolutionary algorithms in concert with an expressive dynamical representation is that we can search a broad space of classifiers. Furthermore, because most of these classifiers will not mirror human thought processes, they are able to capture patterns that humans might not notice, and consequently can be used as a source of novel domain knowledge. For example, by analysing evolved classifiers and their distributions, we have made several insights that may be of interest to clinicians: the differential effect of dominance on diagnostic accuracy, the over-representation of certain patterns of acceleration in the movements of PD patients, and combinations of amplitude and frequency which appear to have diagnostic power.

However, a potential disadvantage of this approach is that we may evolve good classifiers, but have little understanding of how they work. Although biologically-motivated algorithms are often competitive against conventional approaches, they are sometimes criticised for producing black boxes whose internal logic is incomprehensible to domain experts. This is a particularly significant issue for medical diagnosis: whilst black box classifiers can be used to guide or support a diagnosis, an automated diagnosis (for instance, during screening) will only be accepted if the medical practitioner has confidence in the basis of the diagnosis. In this paper, we have addressed this through analysis of the evolved classifiers, showing how relatively simple analytical methods can produce significant insight into their diagnostic behaviour. Nevertheless, the picture is incomplete, particularly in regard of the dynamical classifiers, and an important part of future work will be to develop appropriate methods for ascertaining these classifiers' detailed behaviour. Existing work on rule extraction from neural networks [77], and methods for modelling dynamical systems as finite state automata [78], would be good places to start.

Finally, we have also seen that there is a clear advantage to forming ensembles from diverse classifiers, both in terms of classifier behaviour and classifier model. The use of classifier ensembles is a common theme in the machine learning community, and is typified by standard induction techniques such as boosting and bagging [79]. Recently, there has been a growing interest in how evolutionary algorithms can be used to induce classifier ensembles. An evolutionary algorithm's population is a natural source of diversity, both within and between runs, and methods such as multi-objective ranking [25] and co-evolution [80], [81] can effectively leverage this resource in order to generate effective ensembles. It seems likely that these kinds of techniques could be used to further improve the diagnostic accuracy of PD classifiers, and this is something we plan to look at in future work.

VI. Conclusions

In this paper, we have shown how evolutionary algorithms can be used to induce classifiers able to discriminate Parkinson's disease patients from age-matched controls with accuracies in the region of 95%. The classifiers were trained using acceleration time series data collected whilst subjects carried out a standard clinical finger tapping task. To capture the multi-scale patterns present in this data, we used two different classifier architectures: sliding window genetic programming expressions, and artificial biochemical networks. Behavioural analysis indicated that the induced classifiers were able to capture a diverse range of patterns which discriminate the movements of Parkinson's patients from those of neurologically healthy controls. By forming classifier ensembles, we then showed how behaviourally diverse classifiers provide high discriminative power.

References

[1] C. A. Ross and W. W. Smith, "Gene-environment interactions in Parkinson's disease." *Parkinsonism Relat Disord*, vol. 13 Suppl 3, pp. S309–S315, 2007. [Online]. Available: http://dx.doi.org/10.1016/S1353-8020(08)70022-1

[2] Website of the Parkinson's Disease Society. [Online]. Available: http://www.parkinsons.org.uk

[3] J. J. Jankovic and E. Tolosa, *Parkinson's Disease and Movement Disorders*, 4th ed. Philadelphia: Lippincott Williams and Wilkins, 2002.

[4] S. Fahn and The Parkinson Study Group, "Does levodopa slow or hasten the rate of progression of Parkinson's disease?" *J Neurol*, vol. 252 Suppl 4, pp. IV37–IV42, Oct 2005.

[5] J. Jankovic, A. H. Rajput, M. P. McDermott, and D. P. Perl, "The evolution of diagnosis in early Parkinson disease," *Arch Neurol*, vol. 57, no. 3, pp. 369–372, Mar 2000.

[6] A. Schrag, Y. Ben-Shlomo, and N. Quinn, "How valid is the clinical diagnosis of Parkinson's disease in the community?" *J Neurol Neurosurg Psychiatry*, vol. 73, no. 5, pp. 529–534, Nov 2002.

[7] A. J. Hughes, S. E. Daniel, L. Kilford, and A. J. Lees, "Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases." *J Neurol Neurosurg Psychiatry*, vol. 55, no. 3, pp. 181–184, Mar 1992.

[8] K. L. Possin and D. I. Kaufer, "Parkinsonian dementias." *Continuum (Minneapolis Minn)*, vol. 16, no. 2, Dementia, pp. 57–79, Apr 2010. [Online]. Available: http://dx.doi.org/10.1212/01.CON.0000368212.86835.a8

[9] S. H. Fox, R. Katzenschlager, S.-Y. Lim, B. Ravina, K. Seppi, M. Coelho, W. Poewe, O. Rascol, C. G. Goetz, and C. Sampaio, "The movement disorder society evidence-based medicine review update: Treatments for the motor symptoms of Parkinson's disease." *Mov Disord*, vol. 26 Suppl 3, pp. S2–41, Oct 2011. [Online]. Available: http://dx.doi.org/10.1002/mds.23829

[10] C. G. Goetz, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, G. T. Stebbins, M. B. Stern, B. C. Tilley, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. W. Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. Van Hilten, and N. LaPelle, "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan," *Movement Disorders*, vol. 22, no. 1, pp. 41–47, 2007. [Online]. Available: http://dx.doi.org/10.1002/mds.21198

[11] M. B. Davidson, D. J. M. McGhee, and C. E. Counsell, "Comparison of patient rated treatment response with measured improvement in Parkinson's disease." *J Neurol Neurosurg Psychiatry*, vol. 83, pp. 1001–1005, May 2012. [Online]. Available: http://dx.doi.org/10.1136/jnnp-2012-302741

[12] D. A. Heldman, J. P. Giuffrida, R. Chen, M. Payne, F. Mazzella, A. P. Duker, A. Sahay, S. J. Kim, F. J. Revilla, and A. J. Espay, "The modified bradykinesia rating scale for Parkinson's disease: Reliability and comparison with kinematic measures," *Movement Disorders*, vol. 26, no. 10, pp. 1859–1863, 2011. [Online]. Available: http://dx.doi.org/10.1002/mds.23740

[13] B. Post, M. P. Merkus, R. M. A. de Bie, R. J. de Haan, and J. D. Speelman, "Unified Parkinson's disease rating scale motor examination: are ratings of nurses, residents in neurology, and movement disorders specialists interchangeable?" *Mov Disord*, vol. 20, no. 12, pp. 1577–1584, Dec 2005. [Online]. Available: http://dx.doi.org/10.1002/mds.20640

[14] N. M. Aly, J. R. Playfer, S. L. Smith, and D. M. Halliday, "A novel computer-based technique for the assessment of tremor in Parkinson's disease," *Age and Ageing*, vol. 36, no. 4, pp. 395–399, 2007.

[15] S. L. Smith, P. Gaughan, D. M. Halliday, Q. Ju, N. M. Aly, and J. R. Playfer, "Diagnosis of Parkinson's disease using evolutionary algorithms," *Genetic Programming and Evolvable Machines*, vol. 8, no. 4, pp. 433–447, 2007.

[16] S. L. Smith and J. Timmis, "An immune network inspired evolutionary algorithm for the diagnosis of Parkinson's disease," *BioSystems*, vol. 94, no. 1–2, pp. 34–46, 2008.

[17] A. Ericsson, M. Lonsdale, K. Astrom, L. Edenbrandt, and L. Friberg, "Decision support system for the diagnosis of Parkinson's disease," in *Image Analysis*, ser. Lecture Notes in Computer Science, H. Kalviainen, J. Parkkinen, and A. Kaarna, Eds. Springer Berlin / Heidelberg, 2005, vol. 3540, pp. 275–275.

[18] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease." *IEEE Trans Biomed Eng*, vol. 59, no. 5, pp. 1264–1271, May 2012. [Online]. Available: http://dx.doi.org/10.1109/TBME.2012.2183367

[19] M. A. Lones and S. L. Smith, "Discriminating normal and cancerous thyroid cell lines using implicit context representation Cartesian genetic programming," in *Proc. IEEE Congress on Evolutionary Computation (CEC) 2010*, P. Sobrevilla, Ed. IEEE Press, 2010, pp. 1–6.

[20] ——, "Objective assessment of visuo-spatial ability using implicit context representation Cartesian genetic programming," in *Genetic and Evolutionary Computation: Medical Applications*, S. L. Smith and S. Cagnoni, Eds. Chichester, UK: John Wiley & Sons, 2011, pp. 174–189.

[21] National Institute for Health and Clinical Excellence, *Parkinson's disease: diagnosis and management in primary and secondary care*. London: Royal College of Physicians, 2006. [Online]. Available: http://www.nice.org.uk/CG035

[22] M. Zhang and P. Wong, "Genetic programming for medical classification: a program simplification approach," *Genetic Programming and Evolvable Machines*, vol. 9, no. 3, pp. 229–255, Sep. 2008.

[23] T. Paul and H. Iba, "Prediction of cancer class with majority voting genetic programming classifier using gene expression data," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 6, no. 2, pp. 353 –367, april-june 2009.

[24] S. Winkler, M. Affenzeller, and S. Wagner, "Using enhanced genetic programming techniques for evolving classifiers in the context of medical diagnosis," *Genetic Programming and Evolvable Machines*, vol. 10, no. 2, pp. 111–140. [Online]. Available: http://dx.doi.org/10.1007/s10710-008-9076-8

[25] U. Bhowan, M. Johnston, M. Zhang, and X. Yao, "Evolving diverse ensembles using genetic programming for classification with unbalanced data," *Evolutionary Computation, IEEE Transactions on*, vol. 17, no. 3, pp. 368–386, 2013.

[26] A. A. Freitas, "A review of evolutionary algorithms for data mining," in *Soft Computing for Knowledge Discovery and Data Mining*, O. Maimon and L. Rokach, Eds. Boston, MA: Springer US, 2008, pp. 79–111.

[27] G. B. Fogel, "Computational intelligence approaches for pattern discovery in biological systems," *Briefings in bioinformatics*, vol. 9, no. 4, pp. 307–316, 2008.

[28] S. L. Smith and S. Cagnoni, *Genetic and evolutionary computation: medical applications*. Chichester, UK: John Wiley & Sons, 2011.

[29] J. H. Moore, L. W. Hahn, M. D. Ritchie, T. A. Thornton, and B. C. White, "Routine discovery of complex genetic models using genetic algorithms," *Applied soft computing*, vol. 4, no. 1, pp. 79–86, 2004.

[30] M. A. Lones and A. M. Tyrrell, "The evolutionary computation approach to motif discovery in biological sequences," in *Genetic and Evolutionary Computation Conference, GECCO 2005, Workshop Proceedings*, F. Rothlauf, Ed. New York: ACM, 2005, pp. 1–11.

[31] J. Koza, F. Bennett, and D. Andre, "Using programmatic motifs and genetic programming to classify protein sequences as to cellular location," in *Evolutionary Programming VII, 7th International Conference, EP98*, V. W. Porto, N. Saravanan, D. E. Waagen, and A. E. Eiben, Eds. Berlin/Heidelberg: Springer, 1998, pp. 437–447.

[32] W. Ashlock and S. Datta, "Evolved features for DNA sequence classification and their fitness landscapes," *Evolutionary Computation, IEEE Transactions on*, vol. 17, no. 2, pp. 185–197, 2013.

[33] L. Findley, M. Gresty, and G. Halmagyi, "Tremor, the cogwheel phenomenon and clonus in Parkinson's disease." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 44, no. 6, pp. 534–546, 1981.

[34] P. G. Espejo, S. Ventura, and F. Herrera, "A survey on the application of genetic programming to classification," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 2, pp. 121–144, 2010.

[35] A. T. Harris, A. Lungari, C. J. Needham, S. L. Smith, M. A. Lones, S. E. Fisher, X. B. Yang, N. Cooper, J. Kirkham, D. A. Smith, D. P. Martin-Hirsch, and A. S. High, "Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample," *Head & Neck Oncology*, vol. 1:34, September 2009.

[36] J. Larres, M. Zhang, and W. N. Browne, "Using unrestricted loops in genetic programming for image classification," in *IEEE Congress on Evolutionary Computation (CEC 2010)*, P. Sobrevilla, Ed. Barcelona, Spain: IEEE Press, 2010.

[37] S. Stepney, "Nonclassical computation: a dynamical systems perspective," in *Handbook of Natural Computing*, G. Rozenberg, T. Bäck, and J. N. Kok, Eds. Berlin/Heidelberg: Springer, 2009, vol. 2, ch. 52.

[38] M. A. Lones, L. A. Fuente, A. P. Turner, L. S. D. Caves, S. Stepney, S. L. Smith, and A. M. Tyrrell, "Artificial biochemical networks: Evolving dynamical systems to control dynamical systems," *Evolutionary Computation, IEEE Transactions on*, 2013, in press. [Online]. Available: http://dx.doi.org/10.1109/TEVC.2013.2243732

[39] M. Hüsken and P. Stagge, "Recurrent neural networks for time series classification," *Neurocomputing*, vol. 50, pp. 223–235, 2003.

[40] T. Verplancke, S. Van Looy, K. Steurbaut, D. Benoit, F. De Turck, G. De Moor, and J. Decruyenaere, "A novel time series analysis approach for prediction of dialysis in critically ill patients using echo-state networks," *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 4, 2010.

[41] M. A. Lones, S. L. Smith, A. M. Tyrrell, J. E. Alty, and D. R. S. Jamieson, "Characterising neurological time series data using biologically-motivated networks of coupled discrete maps," *BioSystems*, vol. 112, no. 2, pp. 94–101, 2013.

[42] S. L. Smith, S. Leggett, and A. M. Tyrrell, "An implicit context representation for evolving image processing filters," in *Applications of Evolutionary Computing, Proc. EvoWorkshops 2005*, ser. Lecture Notes in Computer Science, F. Rothlauf, J. Branke, S. Cagnoni, D. W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G. D. Smith, and G. Squillero, Eds., vol. 3449, 2005, pp. 407–416.

[43] M. A. Lones and A. M. Tyrrell, "Biomimetic representation with enzyme genetic programming," *Genetic Programming and Evolvable Machines*, vol. 3, no. 2, pp. 193–217, June 2002.

[44] M. A. Lones, "Enzyme genetic programming: Modelling biological evolvability in genetic programming," Ph.D. dissertation, Department of Electronics, University of York, 2003.

[45] M. A. Lones and A. M. Tyrrell, "Modelling biological evolvability: Implicit context and variation filtering in enzyme genetic programming," *BioSystems*, vol. 76, no. 1–3, pp. 229–238, August–October 2004.

[46] J. F. Miller and P. Thomson, "Cartesian genetic programming," in *Third European Conf. Genetic Programming*, ser. Lecture Notes in Computer Science, R. Poli, W. Banzhaf, W. B. Langdon, J. F. Miller, P. Nordin, and T. C. Fogarty, Eds., vol. 1802, 2000.

[47] X. Cai, S. L. Smith, and A. M. Tyrrell, "Positional independence and recombination in Cartesian genetic programming," in *Genetic Programming, Proc. 9th European Conference, EuroGP 2006*, ser. Lecture Notes in Computer Science, P. Collet, M. Tomassini, M. Ebner, S. Gustafson, and A. Ekárt, Eds., vol. 3905, 2006, pp. 351–360.

[48] W. Langdon, "Size fair and homologous tree genetic programming crossovers," *Genetic Programming and Evolvable Machines*, vol. 1, no. 1/2, pp. 95–119, 2000.

[49] J. Bongard, "A functional crossover operator for genetic programming," in *Genetic Programming Theory and Practice VII*, ser. Genetic and Evo-

lutionary Computation, R. Riolo, U.-M. O'Reilly, and T. McConaghy, Eds. Springer US, 2010, pp. 195–210.

[50] M. A. Lones, A. M. Tyrrell, S. Stepney, and L. S. D. Caves, "Controlling complex dynamics with artificial biochemical networks," in *Genetic Programming, Proc. 13th European Conference, EuroGP 2010*, ser. Lecture Notes in Computer Science, A. I. Esparcia-Alcázar, A. Ekázrt, S. Silva, S. Dignum, and A. Şima Uyar, Eds., vol. 6021. Springer, 2010, pp. 159–170.

[51] ——, "Controlling legged robots with coupled artificial biochemical networks," in *Advances in Artificial Life, ECAL 2011: Proc. 11th European Conference on the Synthesis and Simulation of Living Systems*, T. Lenaerts, M. Giacobini, H. Bersini, P. Bourgine, M. Dorigo, and R. Doursat, Eds. Cambridge, MA: MIT Press, August 2011, pp. 465–472.

[52] A. P. Turner, M. A. Lones, L. A. Fuente, S. Stepney, L. S. D. Caves, and A. Tyrrell, "The artificial epigenetic network," in *Proc. 2013 Symposium Series on Computational Intelligence (SSCI)*. IEEE Press, 2013.

[53] L. A. Fuente, M. A. Lones, A. P. Turner, A. M. Tyrrell, S. Stepney, and L. S. D. Caves, "Computational models of signalling networks for non-linear control," *BioSystems*, vol. 112, no. 2, pp. 122–130, 2013.

[54] K. Kaneko, "Overview of coupled map lattices," *Chaos*, vol. 2, no. 3, pp. 279–282, 1992.

[55] R. M. May, "Simple mathematical models with very complicated dynamics," *Nature*, vol. 261, pp. 459–467, 1976.

[56] B. V. Chirikov, "Research concerning the theory of nonlinear resonance and stochasticity," Institute of Nuclear Physics, Novosibirsk, Tech. Rep. N 267, 1969.

[57] T. Tél and M. Gruiz, *Chaotic Dynamics: An Introduction Based on Classical Mechanics*. Cambridge, UK: Cambridge Press, 2006.

[58] V. Arnold and A. Avez, *Ergodic problems in classical mechanics*. New York: Benjamin, 1968.

[59] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

[60] G. Patterson and M. Zhang, "Fitness functions in genetic programming for classification with unbalanced data," in *AI 2007: Advances in Artificial Intelligence*, ser. Lecture Notes in Computer Science, M. Orgun and J. Thornton, Eds. Springer Berlin Heidelberg, 2007, vol. 4830, pp. 769–775.

[61] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[62] H. C. Kraemer, G. A. Morgan, N. L. Leech, J. A. Gliner, J. J. Vaske, and R. J. Harmon, "Measures of clinical significance." *J Am Acad Child Adolesc Psychiatry*, vol. 42, no. 12, pp. 1524–1529, Dec 2003.

[63] J.-W. Kim, J.-H. Lee, Y. Kwon, C.-S. Kim, G.-M. Eom, S.-B. Koh, D.-Y. Kwon, and K.-W. Park, "Quantification of bradykinesia during clinical finger taps using a gyrosensor in patients with Parkinson's disease." *Med Biol Eng Comput*, vol. 49, no. 3, pp. 365–371, Mar 2011. [Online]. Available: http://dx.doi.org/10.1007/s11517-010-0697-8

[64] H. Ling, L. A. Massey, A. J. Lees, P. Brown, and B. L. Day, "Hypokinesia without decrement distinguishes progressive supranuclear palsy from Parkinson's disease." *Brain*, vol. 135, no. Pt 4, pp. 1141–1153, Apr 2012. [Online]. Available: http://dx.doi.org/10.1093/brain/aws038

[65] D. Halliday, J. Rosenberg, A. Amjad, P. Breeze, B. Conway, and S. Farmer, "A framework for the analysis of mixed time series/point process data — theory and application to the study of physiological tremor, single motor unit discharges and electromyograms," *Progress in Biophysics and Molecular Biology*, vol. 64, no. 2-3, pp. 237 – 278, 1995.

[66] A. H. Rajput, B. Rozdilsky, and L. Ang, "Occurrence of resting tremor in parkinson's disease," *Neurology*, vol. 41, no. 8, 1991.

[67] M. J. Barrett, S. A. Wylie, M. B. Harrison, and G. F. Wooten, "Handedness and motor symptom asymmetry in Parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 82, no. 10, pp. 1122–1124, 2011.

[68] A. van der Hoorn, A. L. Bartels, K. L. Leenders, and B. M. de Jong, "Handedness and dominant side of symptoms in Parkinson's disease." *Parkinsonism Relat Disord*, vol. 17, no. 1, pp. 58–60, Jan 2011. [Online]. Available: http://dx.doi.org/10.1016/j.parkreldis.2010.10.002

[69] H. Chaté and J. Losson, "Non-trivial collective behavior in coupled map lattices: A transfer operator perspective," *Physica D: Nonlinear Phenomena*, vol. 103, no. 1-4, pp. 51 – 72, 1997.

[70] K. Martens and Q. Almeida, "Dissociating between sensory and perceptual deficits in PD: More than simply a motor deficits," *Movement Disorders*, 2012.

[71] Y. Liu, X. Yao, and T. Higuchi, "Evolutionary ensembles with negative correlation learning," *Evolutionary Computation, IEEE Transactions on*, vol. 4, no. 4, pp. 380–387, 2000.

[72] L. M. Bachmann, M. A. Puhan, G. ter Riet, and P. M. Bossuyt, "Sample sizes of studies on diagnostic accuracy: literature survey," *BMJ*, vol. 332, no. 7550, pp. 1127–1129, 5 2006.

[73] W. Chaovalitwongse, R. Pottenger, S. Wang, Y.-J. Fan, and L. Iasemidis, "Pattern- and network-based classification techniques for multichannel medical data signals to improve brain diagnosis," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, no. 5, pp. 977 –988, sept. 2011.

[74] J. V. Marcos, R. Hornero, D. Álvarez, F. del Campo, C. Zamarrón, and M. López, "Utility of multilayer perceptron neural network classifiers in the diagnosis of the obstructive sleep apnoea syndrome from nocturnal oximetry," *Computer Methods and Programs in Biomedicine*, vol. 92, no. 1, pp. 79 – 89, 2008.

[75] L. Guo, D. Rivero, and A. Pazos, "Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks," *Journal of Neuroscience Methods*, vol. 193, no. 1, pp. 156 – 163, 2010.

[76] J. Frank, S. Mannor, and D. Precup, "Activity and gait recognition with time-delay embeddings," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*, M. Fox and D. Poole, Eds. Palo Alto, CA: AAAI Press, 2010.

[77] H. Jacobsson, "Rule extraction from recurrent neural networks: A taxonomy and review," *Neural Computation*, vol. 17, pp. 1223–1263, 2005.

[78] J. P. Crutchfield, "The calculi of emergence: computation, dynamics and induction," *Physica D: Nonlinear Phenomena*, vol. 75, no. 1-3, pp. 11 – 54, 1994.

[79] P. Bühlmann, "Bagging, boosting and ensemble methods," in *Handbook of Computational Statistics*, ser. Springer Handbooks of Computational Statistics, J. E. Gentle, W. K. Härdle, and Y. Mori, Eds. Springer Berlin Heidelberg, 2012, pp. 985–1022.

[80] M. A. Lones and A. M. Tyrrell, "A co-evolutionary framework for regulatory motif discovery," in *2007 IEEE Congress on Evolutionary Computation*, D. Srinivasan and L. Wang, Eds. IEEE Press, 2007, pp. 3894–3901.

[81] R. Thomason and T. Soule, "Novel ways of improving cooperation and performance in ensemble classifiers," in *Proceedings of Genetic and Evolutionary Computation Conference (GECCO 2007)*, D. Thierens, Ed. New York: ACM Press, 2007.

Michael A. Lones (M'01—SM'10) received the M.Eng. degree in computer science in 1999 and the Ph.D. degree in electronics in 2003, both from the University of York, U.K.

In 2004, he received an ERCIM fellowship to carry out research in the Bioinformatics & Gene Regulation group, Faculty of Medicine, Norwegian University of Science and Technology. From 2005 to 2013, he worked in the Intelligent Systems Group, Department of Electronics, University of York, where he was most recently a Lecturer. In 2013, he moved to Heriot-Watt University, Edinburgh to take up a lectureship in the School of Mathematical and Computer Sciences. His current research focuses on biologically motivated models of computation and their application to problems in computational biology, medical informatics, complexity science and robotics.

Dr. Lones is a member of the IEEE Computational Intelligence Society and an active member of the IEEE Technical Committee on Bioinformatics and Bioengineering.

Stephen L. Smith (M'11) received the B.Sc. in computer science in 1984, the M.Sc. degree in electronic engineering in 1986, and the Ph.D. degree in electronic engineering in 1990, all from the University of Kent, Canterbury, U.K.

Since 1994, he has been with the Department of Electronics, University of York, U.K., and is currently a Senior Lecturer. His current main research interests include developing novel representations for evolutionary algorithms, particularly with application to problems in medicine. His work is currently centered on the diagnosis of neurological dysfunction and analysis of mammograms. He has authored over 75 refereed publications.

Dr. Smith is a Chartered Engineer and a Fellow of the British Computer Society. He is an Associate Editor for the journal Genetic Programming and Evolvable Machines and a member of the editorial board for the International Journal of Computers in Healthcare and Neural Computing Applications.

Jane E. Alty received the B.A. degree in medical sciences in 1997 and the M.B. B.Chir. degree in medicine and surgery in 1999 from the University of Cambridge, U.K. In 2003 she received MRCP from the Royal College of Physicians, London.

She has undertaken postgraduate neurology training in England and Australia at The National Hospital for Neurology and Neurosurgery, Leeds Teaching Hospitals NHS Trust, York Teaching Hospitals Foundation Trust and Monash Medical Centre, Melbourne. In 2013, she was appointed a Consultant Neurologist with a specialist interest in movement disorders at Leeds Teaching Hospitals NHS Trust. She is currently undertaking a Doctorate in Medicine at the University of York and her research focuses on measuring bradykinesia and dyskinesia in Parkinson's disease.

Dr. Alty is a member of the Association of British Neurologists, North of England Neurological Association and the Movement Disorders Society.

Stuart E. Lacy (M'12) received an MEng in electronic engineering in 2012 from the University of York, U.K.

Since October 2012 he has been working towards a PhD in electronic engineering, in the Intelligent Systems Research Group at the University of York. His area of research involves applying biologically inspired computational techniques to medical problems.

Mr. Lacy is a Graduate Student Member of the IEEE.

Katherine L. Possin received the Ph.D. degree in clinical neuropsychology from the University of California San Diego in 2007.

Currently, Dr. Possin is an Assistant Professor of Neurology at the University of California San Francisco Memory and Aging Center. Her research examines the neural bases of cognitive functions in neurodegenerative diseases. She aims to develop anatomically-specific cognitive measures to assist with early diagnosis and disease monitoring, and innovative technologies to improve dementia care.

Dr. Possin is a member of the American Academy of Neurology and the International Neuropsychological Society.

D. R. Stuart Jamieson studied biochemistry at Oxford University gaining B.A and M.A. degrees before training in medicine at the University of Birmingham Medical School, U.K.

In 1990, he took up a Medical Research Council Training Fellowship at the MRC Virology Unit, Glasgow obtaining a Ph.D. in 1993. Subsequently he returned to the Institute of Neurological Sciences, Glasgow, to continue his career as a registrar in neurology, later being appointed a clinical lecturer and honorary senior registrar. In 1997, he became a consultant neurologist at the Leeds Teaching Hospitals NHS Trust, and an honorary senior clinical lecturer at the University of Leeds. His current medical research is in the field of movement disorders, especially Parkinson's disease. He is actively involved in research using information technology to diagnose and manage different aspects of neurodegenerative diseases.

Dr. Jamieson is a general neurologist with a specialist interest in movement disorders particularly Parkinson's disease and is a member of the Association of British Neurologists and the Movement Disorders Society.

Andy M. Tyrrell (SM'96) received the first class honors degree in 1982 and the Ph.D. degree in 1985 from Aston University, Birmingham, U.K., both in electrical and electronic engineering.

Since April 1990, he has been with the Department of Electronics, University of York. He was promoted to the Chair of Digital Electronics in 1998. He is Head of the Intelligent Systems research group at York and was Head of Department between 2000 and 2007. His current research interests include the design of biologically inspired architectures, artificial immune systems, evolvable hardware, FPGA system design, parallel systems, fault tolerant design, and real-time systems. He has published over 260 papers in these areas and attracted funds in excess of 6.5M. In particular, over the last 15 years, his research group at the University of York has concentrated on bio-inspired systems.

Prof. Tyrrell is a fellow of the IET.